# Distributed Learning under Imperfect Sensing in Cognitive Radio Networks

Keqin Liu*, Qing Zhao*, Bhaskar Krishnamachari◇
*University of California, Davis, CA, 95616, USA
{kqliu, qzhao}@ucdavis.edu
◇University of Southern California, Los Angeles, CA, 90089, USA
bkrishna@usc.edu

**Abstract**

We consider a cognitive radio network, where $M$ distributed secondary users search for spectrum opportunities among $N$ independent channels without information exchange. The occupancy of each channel by the primary network is modeled as Bernoulli process with unknown mean which represents the unknown traffic load of the primary network. In each slot, a secondary transmitter chooses one channel to sense and subsequently transmit if the channel is sensed as idle. Sensing is considered to be imperfect, *i.e.,* an idle channel can be sensed as busy and vice versa. Users transmit on the same channel collide and none of them can transmit successfully. The objective is to maximize the system throughput under the collision constraint imposed by the primary network while ensuring synchronous channel selection between each secondary transmitter and its receiver. The performance of a channel selection policy is measured by the system regret, defined as the expected total performance loss with respect to the optimal performance under the ideal scenario where all channel means are known to all users and collisions among users are eliminated throughput perfect scheduling. We show that the optimal system regret rate is at the same logarithmic order as the *centralized* counterpart with *perfect sensing*. An order-optimal decentralized policy is constructed to achieve the logarithmic order of the system regret rate while ensuring the fairness among all users.

**Index Terms**

Cognitive radio, distributed learning, regret, decentralized multi-armed bandit, imperfect observation

[0]This work was supported by the Army Research Laboratory under the NS-CTA Grant W911NF-09-2-0053, and by the Army Research Office under Grant W911NF-08-1-0467.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **JUN 2010** | | **00-00-2010 to 00-00-2010** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Distributed Learning under Imperfect Sensing in Cognitive Radio Networks** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Department of Electrical and Computer Engineering,Davis,CA,95616** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

**We consider a cognitive radio network, where M distributed secondary users search for spectrum opportunities among N independent channels without information exchange. The occupancy of each channel by the primary network is modeled as Bernoulli process with unknown mean which represents the unknown traffic load of the primary network. In each slot, a secondary transmitter chooses one channel to sense and subsequently transmit if the channel is sensed as idle. Sensing is considered to be imperfect, i.e., an idle channel can be sensed as busy and vice versa. Users transmit on the same channel collide and none of them can transmit successfully. The objective is to maximize the system throughput under the collision constraint imposed by the primary network while ensuring synchronous channel selection between each secondary transmitter and its receiver. The performance of a channel selection policy is measured by the system regret, defined as the expected total performance loss with respect to the optimal performance under the ideal scenario where all channel means are known to all users and collisions among users are eliminated throughput perfect scheduling. We show that the optimal system regret rate is at the same logarithmic order as the centralized counterpart with perfect sensing. An order-optimal decentralized policy is constructed to achieve the logarithmic order of the system regret rate while ensuring the fairness among all users.**

15. SUBJECT TERMS
**Cognitive radio, distributed learning, regret, decentralized multi-armed bandit, imperfect observation**

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **Same as Report (SAR)** | **18** | |

## I. INTRODUCTION

We consider a distributed learning problem arisen in the context of cognitive radio networks. There are multiple distributed secondary users searching for idle channels temporarily unused by the primary network. We assume that the state—1 (idle) or 0 (busy)—of each channel evolves as an i.i.d. Bernoulli process across time slots with an unknown mean which represents the unknown traffic load of the primary network. At the beginning of each slot, each secondary transmitter chooses one channel to sense and subsequently transmits to its receiver if the channel is sensed as idle. Sensing is subject to errors: an idle channel may be sensed as busy and *vice versa*. If the transmission is successful, the secondary receiver sends back an acknowledgement (ACK) to the transmitter over the same channel at the end of the slot. The secondary users do not exchange information on their decisions and observations. There are two types of collisions that may occur: a *primary* collision happens when a secondary user transmits in a busy channel and a *secondary* collision happens when multiple secondary users transmit in the same channel. In either case, the transmission fails. The objective is to design a decentralized channel selection policy for optimal long-term network throughput under a constraint on the maximum probability of *primary collisions*.

Another important design constraint is the synchronous channel selection between each secondary transmitter and its receiver. We do not assume any dedicated control channel to coordinate each pair of the secondary transmitter and receiver. To ensure synchronization, they can either make the decision based on the common observation history (*i.e.*, number of ACKs observed from each channel) or exploiting the idle channels to exchange control information to coordinate. The tradeoff involved here is that the information from ACKs may not be sufficient for learning the channel rank due to collisions while additional communications between a secondary transmitter and its receiver causes a sacrifice in the throughput.

We measure the performance of a decentralized policy by the system regret, which is defined as the expected total data loss with respect to the optimal performance under the ideal scenario where all channel means are known to all users and collisions among users are eliminated throughput perfect scheduling. The objective is to minimize the rate at which the regret grows with time. Note that the system regret rate is a finer performance measure than the long-term throughput. All policies with a sublinear regret rate would achieve the maximum long-term throughput. However, the difference in their performance measured by the expected total bits of transmitted data over a time horizon of length $T$ can be arbitrarily large as $T$ grows. It is thus of greatly interest to characterize the minimum regret rate and construct

policies optimal under this finer performance measure.

The above problem involves a complicated dilemma of exploitation, exploration, and competition. Specifically, each user needs to learn the channel rank efficiently in order to choose the best channels while avoiding significant collisions to other users. Compared to the scenario of perfect sensing, learning the channel rank under imperfect sensing is substantially more challenging due to the imperfect observation of channel states and the synchronization constraint between each secondary receiver and its transmitter.

In this paper, we show that the minimum system regret rate is at the same logarithmic order as the *centralized* counterpart with *perfect sensing*. A decentralized policy is constructed to achieve this optimal order. Under this policy, the system throughput quickly converges to the maximum throughput in the ideal scenario of known channel model and centralized scheduling. The proposed policy further achieves the fairness among users, *i.e.*, all users converge to the same local throughput at the same rate as time goes to infinity. Last, we extend the problem to general decentralized multi-armed bandits (MAB) with imperfect observation models where control information exchange between the transmitter and the receiver is prohibited. A decentralized policy is proposed to achieve the $O(\sqrt{T})$ regret rate with time $T$.

**Related Work** Under perfect sensing, the cognitive radio network with unknown Bernoulli channel model and multiple distributed users was considered in [1–3]. In [1], a heuristic policy based on histogram estimation of the unknown parameters was proposed. This policy provides a linear order of the system regret rate, thus cannot achieve the maximum throughput. In [2], the problem is formulated as a decentralized MAB, which generalizes the classic MAB with a single user [4,5]. A time division fair sharing (TDFS) framework for constructing order-optimal and fair decentralized policies is proposed under general reward, observation, and collision models. In [3], order-optimal distributed policies were established based on the single-user polices proposed in [6]. Compared to the TDFS policies developed in [2], the policies proposed in [3] are limited to Bernoulli reward models and cannot achieve fairness among users. In [7], a more general channel model that allows each channel to have different means for different users is considered under perfect sensing. A centralized policy that assumes full information exchange and cooperation among users is proposed which achieves the logarithmic order of the regret rate.

**Notation** Let $|\mathcal{A}|$ denote the cardinality of set $\mathcal{A}$. For two positive integers $k$ and $l$, define $k \oslash l \triangleq ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \cdots, l$.

## II. NETWORK MODEL

Consider the spectrum consisting of $N$ independent but nonidentical channels and $M$ distributed secondary users. Each user consists of one transmitter and one receiver. Let $\mathbf{S}(t) = [S_1(t), \cdots, S_N(t)] \in \{0,1\}^N$ $(t \geq 1)$ denote the system state, where $S_i(t) \in \{0 \text{ (busy)}, \ 1 \text{ (idle)}\}$ is the state of channel $i$ in slot $t$ that evolves as as an i.i.d. Bernoulli process with unknown mean $\theta_i \in (0,1)$. We assume that the $M$ largest means are distinct.

In slot $t$, a secondary user (say user $i$ $(1 \leq i \leq M)$) chooses a sensing action $a_i(t) \in \{1, \cdots, N\}$ that specifies the channel (say, channel $n$) to sense based on its observation and decision history. Based on the sensed signals, the user detects the channel state, which can be considered as a binary hypothesis test:

$$\mathcal{H}_0: \ S_n(t) = 1 \text{ (idle)} \ \ vs. \ \ \mathcal{H}_1: \ S_n(t) = 0 \text{ (busy)}.$$

The performance of channel state detection is characterized by the receiver operating characteristics (ROC) which relates the probability of false alarm $\epsilon$ to the probability of miss detection $\delta$:

$$\epsilon \triangleq \Pr\{\text{decide } \mathcal{H}_1 | \mathcal{H}_0 \text{ is true}\}, \quad \delta \triangleq \Pr\{\text{decide } \mathcal{H}_0 | \mathcal{H}_1 \text{ is true}\}.$$

If the detection outcome is $\mathcal{H}_0$, the user accesses the channel for data transmission. The design should be subject to a constraint on the probability of accessing a busy channel, which causes interference to the primary network. Specifically, the probability of collision $\mathcal{P}_n(t)$ perceived by the primary network in any channel and slot is capped below a predetermined threshold $\zeta$, *i.e.,*

$$\mathcal{P}_n(t) \triangleq \Pr(\text{decide } \mathcal{H}_1 | S_n(t) = 0) = \delta \leq \zeta, \quad \forall \ n, \ t.$$

We should set the miss detection probability $\delta = \zeta$ as the detector operating point to minimize the false alarm probability $\epsilon$. If multiple users decide to transmit over the same channel, they collide and no one can transmit successfully. In other words, a secondary user can transmit data successfully if and only if the chosen channel is idle, detected correctly, and no collision happens. Since failed transmissions may occur, acknowledgements (ACKs) are necessary to ensure guaranteed delivery. Specifically, when the receiver successfully receives a packet from a channel, it sends an acknowledgement to the transmitter over the same channel at the end of the slot. Otherwise, the receiver does nothing, *i.e.,* a NAK is defined as the absence of an ACK. We assume that acknowledgements are received without error since acknowledgements are always transmitted over idle channels.

## III. PROBLEM FORMULATION

In each slot, each secondary transmitter and its receiver need to select the same channel for data transmission without a dedicated control channel. One natural way is that the transmitter and its receiver use the common local observation history (ACKs/NAKs) in learning and decision making. However, due to the collisions among secondary users, the information included in previously observed ACKs/NAKs may not be sufficient to learn the unknown channel model efficiently. An alternative approach is to let each transmitter decide whether or not to send its receiver the control information (instead of the objective data) in the chosen channel for future synchronization. We consider the worst scenario that each transmission of the control information occupies an entire idle slot[1]. Since sending the control information causes a sacrifice in the immediate throughput, it should be avoided as much as possible in order to maximize the number of opportunities for transmitting the objective data.

We define a local policy $\pi_i$ for user $i$ as a sequence of functions $\pi_i = \{\pi_i(t)\}_{t\geq 1}$, where $\pi_i(t)$ maps user $i$'s local information that is common to its transmitter and receiver to the sensing action $a_i(t)$ in slot $t$. The decentralized policy $\pi$ is thus given by the concatenation of the local policy for each user: $\pi = [\pi_1, \cdots, \pi_M]$. Define immediate reward $Y(t)$ as the total number of successful transmissions of the objective data by all users in slot $t$:

$$Y(t) = \Sigma_{j=1}^{N} \mathbb{I}'_j(t) S_j(t),$$

where $\mathbb{I}'_j(t)$ is the indicator function that equals to 1 if channel $j$ is accessed by only one user and used for transmitting the objective data, and 0 otherwise.

Let $\Theta = (\theta_1, \theta_2, \cdots, \theta_N)$ be the unknown parameter set and $\sigma$ a permutation such that $\theta_{\sigma(1)} \geq \theta_{\sigma(2)} \geq \cdots \geq \theta_{\sigma(N)}$. The performance measure of a decentralized policy $\pi$ is defined as the system regret

$$R_T^{\pi}(\Theta) = T\Sigma_{j=1}^{M}(1-\epsilon)\theta_{\sigma(j)} - \mathbb{E}_{\pi}[\Sigma_{t=1}^{T} Y(t)].$$

It is easy to see that $T\Sigma_{j=1}^{M}(1-\epsilon)\theta_{\sigma(j)}$ is the maximum expected total reward over $T$ slots under the ideal scenario that the parameter set $\Theta = (\theta_1, \cdots, \theta_N)$ is known and users are centralized.

Note that the regret is always growing with time since users can never identify the channel parameters perfectly. The objective is to minimize the rate at which $R_T(\Theta)$ grows with time $T$ under any parameter set $\Theta$ by choosing the optimal decentralized policy $\pi^*$.

[1]The results established in this paper can be directly extended to a more relaxed *piggybacking* scenario that assumes that the control information occupies negligible capacity and is included in the data package in each slot.

## IV. Optimal Order of the System Regret

In this section, we show that the minimum system regret rate is at the logarithmic order with time, which implies that the system can achieve the maximum throughput at a significantly fast rate.

*Theorem 1:* The optimal order of the system regret rate is logarithmic with time, *i.e.,* for an optimal decentralized policy $\pi^*$, we have, $\forall\ \Theta$,

$$L(\Theta) = \liminf_{T\to\infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} \leq \limsup_{T\to\infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} = U(\Theta) \tag{1}$$

for some constants $L(\Theta)$ and $U(\Theta)$ that depend on $\Theta$.

*Proof:* To prove the lower bound, we consider a genie-aided system where users are centralized and the synchronization constraint on each secondary transmitter and its receiver is removed from consideration. Note that the channel parameters remain unknown to all users in the genie-aided system. It is easy to see that the problem is equivalent to the one with a single user that can sense $M$ channels simultaneously in each slot. For simplicity, we focus on the latter one. In each slot, the user obtains two types of observations from each chosen channel: the detection outcome and the ACK/NAK. In Lemma 1, we show that the regret rate in the genie-aided system is at least logarithmic with time. The proof is thus completed by noticing that the minimum regret rate in the problem at hand is lower bounded by the one in the genie aided system.

*Lemma 1:* Let $\tilde{R}_T^{\pi}(\Theta)$ denote the regret under a policy $\pi$ in the genie-aided system over $T$ slots. If $\tilde{R}_T^{\pi}(\Theta) = o(T^c)\ \forall\ c > 0$ and $\forall\ \Theta$, then, $\forall\ \Theta$,

$$\liminf_{T\to\infty} \frac{\tilde{R}_T^{\pi}(\Theta)}{\log T} \geq (1-\epsilon)\Sigma_{n:\ \mu(\theta_n)<\mu(\theta_{\sigma(M)})} \frac{\mu(\theta_{\sigma(M)}) - \mu(\theta_n)}{G(\theta_n, \theta_{\sigma(M)})}, \tag{2}$$

where

$$G(\theta_i,\theta_j) = (\epsilon\theta_i + (1-\delta)(1-\theta_i)) \log \frac{\epsilon\theta_i + (1-\delta)(1-\theta_i)}{\epsilon\theta_j + (1-\delta)(1-\theta_j)} + \delta(1-\theta_i) \log \frac{\delta(1-\theta_i)}{\delta(1-\theta_j)} + (1-\epsilon)\theta_i \log \frac{(1-\epsilon)\theta_i}{(1-\epsilon)\theta_j}$$

is the K-L distance between two joint distributions of the detection outcome and the ACK/NAK parameterized by $\theta_i$ and $\theta_j$, respectively.

*Proof:* The proof follows a similar line to that of Theorem 3.1 in [5] by combining the detection outcome and ACK/NAK as a single observation vector of an arm. ■

For the upper bound, we show that their exists a decentralized policy that achieves the logarithmic order of the growth rate of the system regret. See Sec. V for details. ■

## V. The Order-Optimal Decentralized Policy

In this section, we establish an order-optimal and fair decentralized policy $\pi_F^*$ to achieve the optimal logarithmic order of the system regret rate. The general structure of the policy is based on the time division fair sharing (TDFS) of the $M$ best channels among the $M$ distributed users. The TDFS structure is first proposed in [2] in the scenario of perfect sensing. Due to the imperfect observation of channel state and the synchronization constraint, extending the TDFS framework to the problem at hand is highly nontrivial.

Specifically, the local policy of each user consists of disjoint rounds of playing the $M$ channels considered to be the best. Different users have different offsets in sensing the sets of $M$ channels. Consider, for example, user $1$ has offset $0$. In each round, the user successively senses the best, second best, $\cdots$, and the $M$th best channels it considers to be. The offset in each user's round-robin schedule can be predetermined (*e.g.,* based on the user's ID).

To achieve the optimal order of the system regret rate, it is crucial that each user efficiently learns and senses the $M$ best channels in the correct order while ensuring the synchronization between each transmitter and its receiver without significant communication overhead. We first propose a synchronization mechanism for each transmitter and its receiver. Based on the symmetry among users, it is sufficient to consider one user, say, user 1. We assume that its transmitter and receiver have a simple initial setup for synchronization, *e.g.,* in the first round, they will both tune to channel $1$, $2$, $\cdots$, $M$ (*i.e.,* the initial channel rank of the $M$ channels considered to be the best is $(1, 2, \cdots, M)$). As shown in Fig. 1, if an ACK is observed, the transmitter will update the channel rank according to its sensing and detection history. If the updated channel rank is different from the current one, the transmitter will keep sending its receiver the updated channel rank (instead of the objective data) until the channel is successfully received (*i.e.,* a new ACK is observed). For simplicity of presentation, we assume that the channel capacity is enough to send the channel rank in one slot when it is idle[2]. Based upon a successful reception of the updated channel rank, the transmitter and receiver will use this new channel rank for channel sensing in the next round. If there is no new channel rank received, they will keep using the previous one. We point out that each round the transmitter only updates the channel rank once based on the first ACK (if exists) received in this round.

Next, we consider the learning of channel rank at the transmitter whenever an update is required. The

[2]Note that the channel rank consists of integer values and only needs finite capacity to transmit. If the channel capacity is not enough to send the channel rank in one slot, the transmitter will send the channel rank in multiple slots.

basic approach is to reducing the problem to the one with the perfect observation model as considered in [2]. Note that the transmitter only uses the detection outcome (not ACKs/NAKs) to learn the channel order at each update. Since the mean of the detection outcome from a channel (say, channel $n$) is equal to $(1-\epsilon)\theta_n$, the channel rank ordered by their state means is the same as that ordered by their detection means. We can thus treat the detection outcome as the *new state* of each channel in learning the channel rank. Consequently, the observation of the new state becomes perfect. The user then adopts a procedure analogous to that in [2] to identify the set of the $M$ best channels. Basically, the user first identifies the best channel by applying a single-user policy (say, Lai-Robbins policy established in [4]) for the classic MAB. To identify the $k$th ($1 \leq k \leq M$) best channel, the user removes the $k-1$ channels considered to have a higher rank than other channels and apply Lai-Robbins policy to the remaining $N-k+1$ channels. The main difference here to that in [2] is that the user needs to identify the entire rank of the $M$ best channels in one shot (as the first ACK is observed in the current round) and the channel sensing under this rank can not be realized until the round before which the receiver has successfully received this rank information and no newer update has been received. Establishing the efficiency for learning the channel rank is thus more challenging compared to the scenario addressed in [2].

A detailed implementation of the decentralized policy $\pi_F^*$ is given in Fig. 2.

*Theorem 2:* Under the decentralized policy $\pi_F^*$, we have

$$\limsup_{T\to\infty} \frac{R_T^{\pi_F^*}(\Theta)}{\log T} = C(\Theta) \tag{3}$$

for some constant $C(\Theta)$ that depends on $\Theta$.

*Proof:* Note that the set of slots in which a reward loss occurs is a subset of slots in which there exist a user that does not sense the correct channel or a transmitter that sends the channel rank information instead of the objective data. It is thus sufficient to prove the expected number of slots that a user does not sense the $M$ best channels in a correct order or its transmitter sends the channel rank information to the receiver is at most logarithmic with time. Without loss of generality, consider user 1. We first present the following lemma, which shows that the expected number of times that the transmitter does not update the channel rank correctly is at most logarithmic with time.

*Lemma 2:* Let $\bar{\tau}_u(T)$ denote the number of times that the channel rank is updated incorrectly at the transmitter, we have

$$\limsup_{T\to\infty} \frac{\bar{\tau}_u(T)}{\log T} = V(\Theta) \tag{4}$$

for some constant $V(\Theta)$ that depend on $\Theta$.

*Proof:* See Appendix A for details. ■

Now we show that the expected number of rounds that a user does not sense the $M$ best channels in a correct order is at most logarithmic with time. Note that expected number of slots between two successive updates at the transmitter is uniformly bounded by some constant. So the expected number of successive rounds that the user does not sense the $M$ best channel in the correct order caused by the previous incorrect update is uniformly bounded by some constant. The expected number of rounds that the user does not sense the $M$ best channels in a correct order is thus at the same order as that of the incorrect updates on the channel rank at the transmitter, which is at most logarithmic with time based on Lemma 2.

Note that the transmitter only needs to send its receiver an update if the the new channel rank is different from the current one. Except that the channel rank is updated incorrectly, the updated channel ranks are all the same. By noticing that the expected number of times that the channel rank is updated incorrectly is at most logarithmic with time, the expected number of times that the transmitter needs to send its receiver the updated channel rank information is at most logarithmic with time. Since each sending duration till a successful reception is uniformly bounded in expectation, the expected number of slots that the transmitter sends its receiver the updated channel rank information is at most logarithmic with time.

We thus proved Theorem 2.

■

Based on the symmetry among users' local policies, we show that $\pi_F^*$ achieves the fairness among all users.

*Theorem 3:* Define the local regret for user $i$ under $\pi_F^*$ as

$$R^{\pi^*_{F,i}}(\Theta) \triangleq \frac{1}{M} T \Sigma_{j=1}^{M} (1-\epsilon)\mu(\theta_{\sigma(j)}) - \mathbb{E}_{\pi^*_F}[\Sigma_{t=1}^{T} Y_i(t)],$$

where $Y_i(t)$ is the immediate reward obtained by user $i$ in slot $t$. We have

$$\limsup_{T\to\infty} \frac{R^{\pi^*_{F,i}}(\Theta)}{\log T} = \frac{1}{M} \limsup_{T\to\infty} \frac{R_T^{\pi^*_F}(\Theta)}{\log T} \ \forall \ i \in \{1, \cdots, M\}.$$

## VI. Extension to General Decentralized MAB with Imperfect Observation Models

In this section, we formulate the decentralized MAB with imperfect observation models that generalizes the cognitive radio problem considered in previous sections.

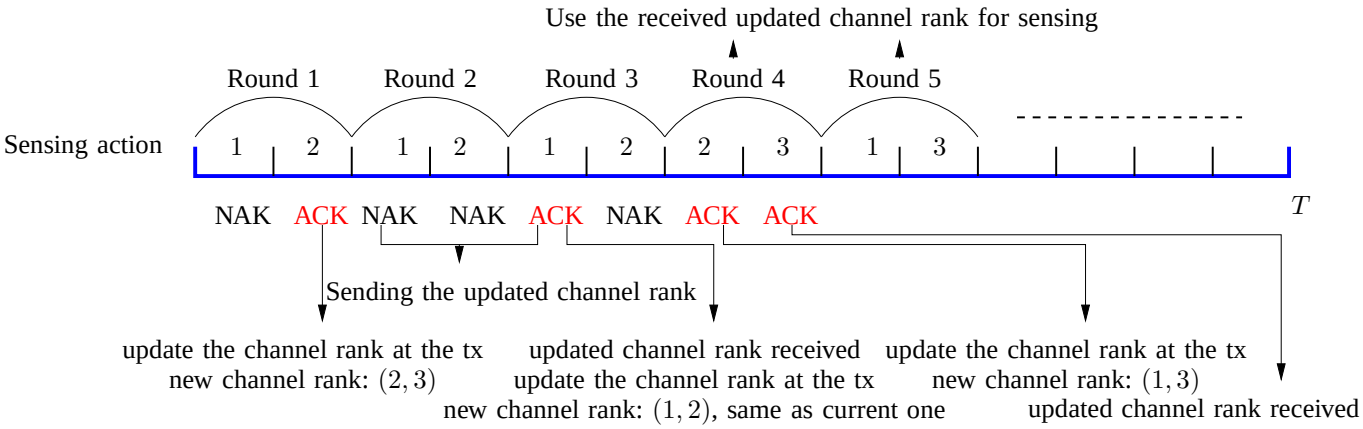


Fig. 1. An example of the structure of user 1's local policy under $\pi_F^*$ ($M = 2$, $N = 3$, tx: transmitter).

In general, there exist $M$ distributed players (users) and $N$ arms (channels) in the system. The reward that each arm can offer is an i.i.d. process with unknown mean. In each slot, each player decides to play one arm based on its local observation and decision history. If multiple players choose the same arm to play, the reward obtained and observed by each of them will be distorted in an arbitrary way (either deterministically or statistically). The cognitive radio problem considered in previous sections can be considered as a special case of the general model, where sensing a channel corresponds to playing an arm and the reward on each arm is given by its state. We point out that, in general, there is no 'transmitter' that can 'sense' the arm state without being affected by collisions.

To design an optimal decentralized policy under the general imperfect observation model, the local observation history of each user needs to be filtered to extract trustable information for learning the arm rank. This could involve a complicated change detection problem and the minimum system regret rate may not achieve the logarithmic order. In this section, we propose a simple policy $\pi_F^g$ to achieve the $O(\sqrt{T})$ regret rate with time $T$ while ensuring the fairness among all players. The following assumptions will be adopted.

A1. The means of the $M$ best arms are nonnegative and distinct.

A2. The variance of the reward from each arm is finite.

The basic idea in $\pi_F^g$ is to constructing a deterministic sequence in which the collisions among players are perfectly avoided. In this sequence, each user plays each of the $N$ arms in a round robin fashion with a different offset. Each user computes the sample mean of each arm solely based on the reward obtained in the slots that belong to this sequence. In other slots that do not belong to this sequence,

**The Decentralized Policy $\pi_F^*$**

Without loss of generality, consider user $i$.

- Notations and Inputs: let $\tilde{\theta}_n(t)$ denote the detection mean obtained from channel $n$ at the transmiiter and $\tau_{n,t}$ the number of times that channel $n$ is sensed up to (but excluding) slot $t$. Let $I(\theta,\theta') = \theta\log(\theta/\theta') + (1-\theta)\log((1-\theta)/(1-\theta'))$ denote the K-L distance between the Bernoulli distributions parameterized by $\theta$ and $\theta'$, respectively. User $i$ first senses each channel once in the first $N$ slots to establish initial observations. Starting from slot $t+1$, user $i$'s local policy consists of disjoint rounds of sensing the $M$ channels considered to be the best. Let $\mathcal{Q}_k$ denote the channel sensing order in the $k$th round. Let $\mathcal{U}_k$ denote the number of updates of channel rank at the transmitter up to (and including) round $k$. Initially, $\mathcal{Q}_1 = (1,2,\cdots,M)$ and $\mathcal{U}_0 = 0$. Select a $b$ $(0 < b < 1/N)$.

- In the $k$th round, user $i$ does the following.

  1. Both the transmitter and receiver sense the channels considered to be the $M$ best in turn according to $\mathcal{Q}_k$. If an ACK is observed and this is the first ACK observed in this round, the transmitter set $\mathcal{U}_k = \mathcal{U}_{k-1} + 1$ and updates the rank of the $M$ channels considered to be the best according to step 2. The transmitter sends the receiver the updated channel rank if it is different from $\mathcal{Q}_k$ until the next ACK observed. If the receiver received a packet consisting of the updated channel rank previously sent by the transmitter, the receiver sends back an ACK and both the transmitter and receiver set $\mathcal{Q}_{k+1}$ equal to the updated channel rank; otherwise $\mathcal{Q}_{k+1} = \mathcal{Q}_k$.
  2. First, the transmitter identifies the best channel. Let $t$ denote the current time. The user chooses between a leader $l_t$ and a round-robin candidate $r_t = \mathcal{U}_k \oslash N$, where the leader $l_t$ is the channel with the largest detection mean among all channels that have been sensed for at least $(\mathcal{U}_k - 1)b$ times. The user chooses the leader $l_t$ as the best if $\tilde{\theta}_{l_t}(t) > \tilde{\theta}_{r_t}(t)$ and $I(\tilde{\theta}_{r_t}(t), \tilde{\theta}_{l_t}(t)) > \log(t-1)/\tau_{r_t,t}$; otherwise the user chooses the round-robin candidate $r_t$ as the best. To identify the $k$th ($k > 1$) best channel, the user removes the set of $k-1$ channels considered to have a higher rank than others from the channel set and then chooses between a leader and a round-robin candidate defined within the remaining channels. Specifically, let $m(t)$ denote the number of times that the same set of the $k-1$ channels is removed. Among all channels that have been sensed for at least $(m(t)-1)b$ times, let $l_t$ denote the leader with the largest detection mean. Let $r_t = m(t) \oslash (N-k+1)$ be the round-robin candidate where, for simplicity, we have assumed that the remaining channels are indexed by $1,\cdots,N-k+1$. The user chooses the leader $l_t$ as the $k$th best if $\tilde{\theta}_{l_t}(t) > \tilde{\theta}_{r_t}(t)$ and $I(\tilde{\theta}_{r_t}(t), \tilde{\theta}_{l_t}(t)) > \log(t-1)/\tau_{r_t,t}$; otherwise the user chooses the round-robin candidate $r_t$ as the $k$th best.

Fig. 2. The decentralized policy $\pi_F^*$.

each user plays the $M$ arms that have the largest sample means in a round robin fashion with a different offset. Since each user is obligated to play each of the $N$ arms in the deterministic sequence. The system regret rate is at least at the order of the number of slots in this sequence as time grows. In other words, the density of the sequence should be small enough. However, the density of the sequence should also be large enough to ensure an efficient learning of the arm rank. This tradeoff between exploitation and exploration needs to be properly addressed in the policy design. We show that by choosing a sequence of which the cardinality grows at the order $O(\sqrt{T})$ with time $T$, the system regret rate can achieve the same growth rate of its cardinality.

A detailed implementation of $\pi_F^g$ is given in Fig. 3.

*Theorem 4:* For the general decentralized MAB with imperfect observation models, the system regret rate under $\pi_F^g$ is at the order $O(\sqrt{T})$. Furthermore, $\pi_F^g$ achieves the fairness among all users, *i.e.*, the local regret rate of each user is the same.

*Proof:* Let $D(t)$ denote the number of slots in the deterministic sequence up to (but excluding) time $t$. Let $\tilde{\theta}_n(t)$ denote the sample mean of channel $n$ based on the observations in the deterministic sequence up to (but excluding) time $t$. From [4], for i.i.d. random variables $\{Y_1, Y_2, \cdots\}$ with a finite variance, $\Pr(|E(Y_1) - (\Sigma_{i=1}^k Y_i)/k| > \epsilon) = o(k^{-1}),\ \forall\ \epsilon > 0$. Choose $0 < \epsilon < \min\{\theta_i - \theta_j :\ 1 \le i < j \le N,\ \theta_i \ne \theta_j\}/2$. We thus have,

$$\begin{aligned} R_T^{\pi_F^g}(\Theta) &\le \Sigma_{t=1}^T O(\Sigma_{i=1}^N \Pr(|\theta_i - \tilde{\theta}_n(t)| > \epsilon)) + O(D(T)) && (5)\\ &= \Sigma_{t=1}^T o(1/D(t)) + O(D(T)) && (6)\\ &= \Sigma_{t=1}^T o(1/t^{1/2}) + O(T^{1/2}). && (7) \end{aligned}$$

Note that

$$\Sigma_{t=1}^T o(1/t^{1/2}) = o(\int_{t=1}^T t^{1/2} dt) = o(T^{1/2}). \quad (8)$$

From (5) and (8), $R_T^{\pi_F^g}(\Theta) = O(D(T)) = O(T^{1/2})$. ■

## VII. Simulation Examples

In this section, we study the performance of the order optimal policy $\pi_F^*$ for the cognitive radio problem and the policy $\pi_F^g$ for the general decentralized MAB with imperfect observation models.

### A. The Performance of $\pi_F^*$ for the Cognitive Radio Network

We consider the scenario that both the channel noise and the signal of the primary network are white Gaussian processes with zero mean but different power densities. The energy detector is adopted that

**The Decentralized Policy $\pi_F^g$**

Without loss of generality, consider user $i$.

- Notations and Inputs: Let $\mathcal{A}(t)$ denote the index set of slots that belong to the deterministic sequence up to (and including) time $t$. Let $D(t) = |\mathcal{A}(t)|$. Initially, $\mathcal{A}(1) = \{1\}$ and $D(1) = 1$. Select an $a > 0$.

- At time $t$, user $i$ does the following.
  1. If $at^{1/2} > D(t)$ and $t \notin \mathcal{A}(t)$, include $t$ in $\mathcal{A}(t)$, set $D(t) = D(t-1)+1$, and then go to step 2. For the case that $at^{1/2} \leq D(t)$, go to step 2 if $t \in \mathcal{A}(t)$ and step 3 otherwise.
  2. User $i$ plays the $((D(t)+i-1) \oslash N)$th arm and update the sample mean of this arm.
  3. User $i$ plays the arm with the $((t+i-1) \oslash M)$th largest sample mean.

Fig. 3. The decentralized policy $\pi_F^g$.

is optimal under the Neyman-Pearson criterion [8]. In Fig. 4, we observe that the regret converges quickly as time goes. In Fig. 5, we plot the constant of the logarithmic order as a function of $N$. We observe that, from this example, the system performs better for smaller detection errors. Furthermore, the system performance is not monotonic as the number of channels increases. This is due to the tradeoff that as $N$ increases, users are less likely to collide but learning the $M$ best channels becomes more difficult. In Fig. 6, we plot the constant of the logarithmic order as a function of $M$. We observe that the system performance degrades as $M$ increases. This is due to the increased competitions and learning load encountered by all users.

### *B. The Performance of $\pi_F^g$ for the General Model*

W compare the performance of $\pi_F^g$ by setting different values of parameter $a$ (see Fig. 3), which equals to the constant of the $O(\sqrt{T})$ order of the cardinality of the deterministic sequence. Each arm has a Bernoulli reward distribution. Intuitively, we want to choose a small $a$ since the regret rate is equal to rate of the cardinality of the sequence. However, from Fig. 7, we observe that the regret under the smaller parameter converges at a much slower rate than that under the larger parameter. This is due to the fact that for any arm, the convergence of the sample mean to the true mean is not fast enough in terms of the number of samples. It is thus better to choose a fairly large parameter when considering the short-horizon performance.

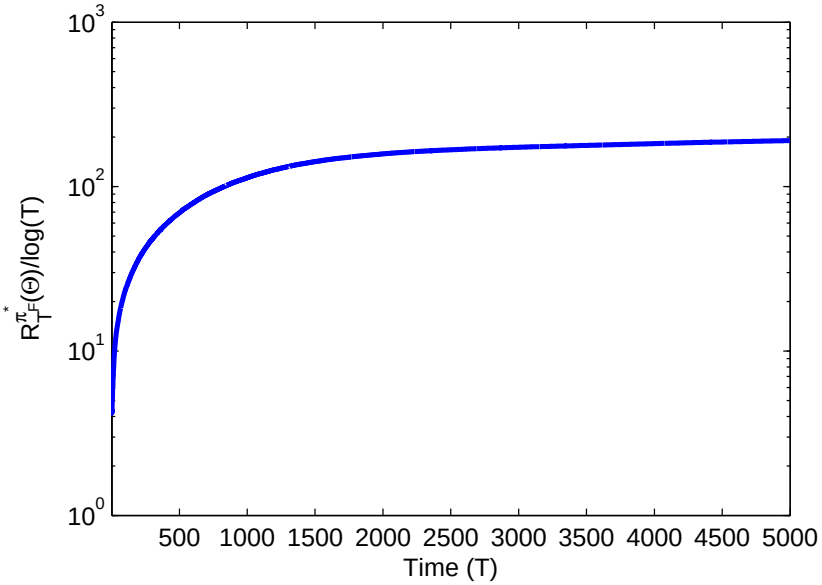


Fig. 4. The Convergence of the regret ($M = 2,\ N = 9,\ \Theta = [0.1,\ 0.2,\ \cdots,\ 0.9]$, $\epsilon = 0.0854$, $\delta = 0.1$, (primary) signal to noise ratio=5db).

## VIII. Conclusion

In this paper, we formulated the distributed learning problem in cognitive radio networks under imperfect sensing. The optimal system regret rate is shown to be at the logarithmic order. An order-optimal decentralized policy is proposed to achieve the logarithmic order of the regret rate and thus lead to a fast convergence to the maximum throughput in the ideal scenario of known channel model and centralized users. Furthermore, the cognitive radio example is extended to the general decentralized MAB with imperfect observation models. A simple decentralized policy is proposed under this general model to achieve the $O(\sqrt{T})$ order of the system regret rate as $T \to \infty$.

## Appendix A. Proof of Lemma 2

We prove by induction on selecting the $M$ best channels. Specifically, it is sufficient to show that, given that the $(i-1)$th best channels are correctly selected, the expected number of updates that the $i$th best channel is not selected correctly is at most logarithmic with time for all $1 \leq i \leq M$.

Let $K$ denote the total number of updates over the horizon of $T$ slots. Let $\mathcal{D}(K)$ denote the set of updates at which the $(i-1)$th best channels are correctly selected up to the $K$th update. For any $\alpha \in (0, \mu(\theta_{\sigma(i)}) - \mu(\theta_{\sigma(i+1)}))$, let $N_1(K)$ denote the number of updates in $\mathcal{D}(K)$ at which channel $n$ is

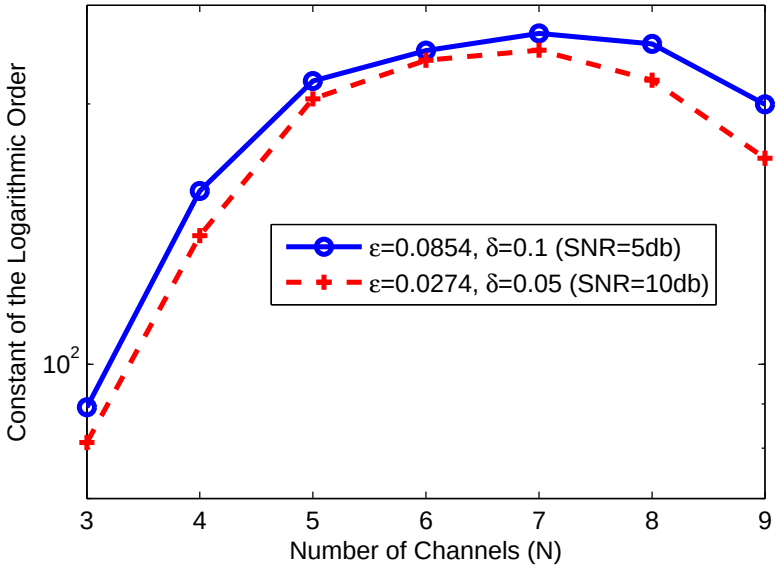


Fig. 5. The performance of $\pi_F^*$ ($T = 5000,\ M = 2,\ \Theta = [0.1,\ 0.2,\ \cdots,\ \frac{N}{10}]$, SNR: (primary) signal to noise ratio).

selected as the $i$th best when $l_t = \sigma(i)$ and $|\tilde{\theta}_{l_t}(t) - (1-\epsilon)\theta_{l_t}(t)| \leq \alpha$ ($t$ is the update time), $N_2(K)$ the number of updates in $\mathcal{D}(K)$ at which channel $n$ is selected as the $i$th best when $l_t = \sigma(i)$ and $|\tilde{\theta}_{l_t}(t) - (1-\epsilon)\theta_{l_t}(t)| > \alpha$, and $N_3(K)$ the number of updates in $\mathcal{D}(K)$ when $l_t \neq \sigma(i)$. It is sufficient to show that $\mathbb{E}[N_1(K)]$, $\mathbb{E}[N_2(K)]$, and $\mathbb{E}[N_3(K)]$ are all at most in the order of $\log T$.

Consider first $\mathbb{E}[N_1(T)]$. We have

$$\begin{aligned}
\mathbb{E}[N_1(k)] &= O(\mathbb{E}[|\{1 \leq k \leq K:\ k \in \{\mathcal{D}(K)\},\ \theta_{l_t} = \theta_{\sigma(i)},\ |\tilde{\theta}_{l_t}(t) - (1-\epsilon)\theta_{l_t}(t)| \leq \alpha,\ \text{and the } k\text{th update is realized}\}|]) \\
&\leq O(\mathbb{E}[|\{1 \leq j \leq T-1:\ \tilde{\theta}_{n\ (j \text{ samples})} \geq \theta_{\sigma(i)} - \alpha \text{ or } I(\tilde{\theta}_{n\ (j \text{ samples})}, \theta_{\sigma(i)} - \alpha) \leq \log(T-1)/j\}|]) \\
&\leq O(\log T), \qquad (9)
\end{aligned}$$

where the first equality is due to the fact that the probability that each update will be realized for channel sensing is lower bounded by some constant non-zero probability, the first inequality is due to the structure of the local policy of $\pi_F^*$, and the second inequality follows the property of Bernoulli distributions established in [4].

Consider $\mathbb{E}[N_2(K)]$. Since the number of observations obtained from $l_t$ at the $s$th ($\forall\ 1 \leq s \leq T$)

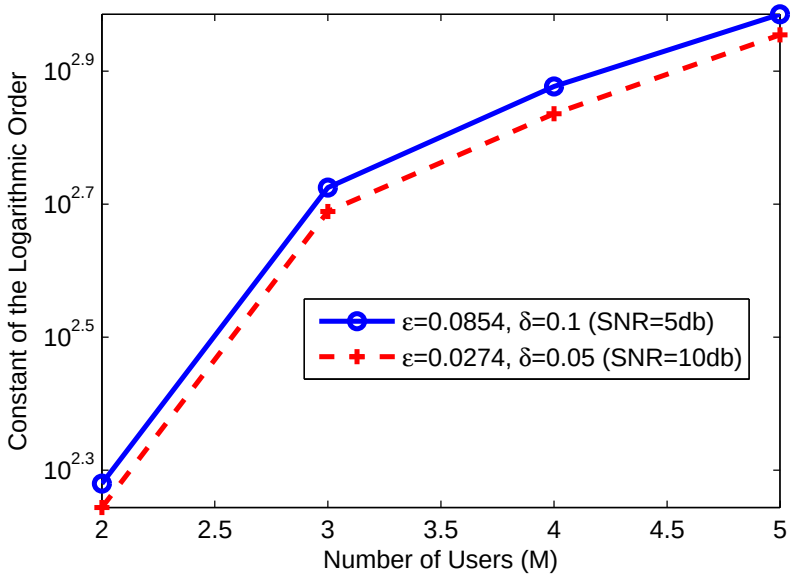


Fig. 6. The performance of $\pi_F^*$ ($T = 5000,\ N = 9,\ \Theta = [0.1,\ 0.2,\ \cdots,\ 0.9]$, SNR: (primary) signal to noise ratio).

update is at least $(s-1)b$, we have that, $\forall\ 1 \le s \le T$,

$$\begin{aligned} & \Pr\{\text{at the } s\text{th update},\ \theta_{l_t} = \theta_{\sigma(i)},\ |\tilde{\theta}_{l_t}(t) - (1-\epsilon)\theta_{l_t}(t)| > \alpha\} \\ \le\ & \Pr\{\sup_{j \ge b(s-1)} |\tilde{\theta}_{l_t}\ (j \text{ samples}) - (1-\epsilon)\theta_{l_t}(t)| > \alpha\} \\ =\ & \Sigma_{i=0}^{\infty} b^i o(s^{-1}) \\ =\ & o(s^{-1}), \end{aligned} \tag{10}$$

where the first equality is due to the property of Bernoulli distributions established in [4].

We thus have,

$$\begin{aligned} \mathbb{E}[N_2(K)] &= \mathbb{E}(|\{1 \le k \le K :\ k \in \mathcal{D}(K),\ \theta_{l_t} = \theta_{\sigma(i)},\ |\tilde{\theta}_{l_t}(t) - (1-\epsilon)\theta_{l_t}(t)| > \alpha\}|) \\ &\le \Sigma_{s=1}^{T} \Pr\{\text{at the } s\text{th update},\ \theta_{l_t} = \theta_{\sigma(i)},\ |\tilde{\theta}_{l_t}(t) - (1-\epsilon)\theta_{l_t}(t)| > \alpha\} \\ &= o(\log T). \end{aligned} \tag{11}$$

Next, we show that $\mathbb{E}[N_3(K)] = o(\log T)$.

Choose $0 < \alpha_1 < (\mu(\theta_{\sigma(i)}) - \mu(\theta_{\sigma(i+1)}))/2$ and $c > (1 - Nb)^{-1}$. For $r = 0, 1, \cdots$, define the following

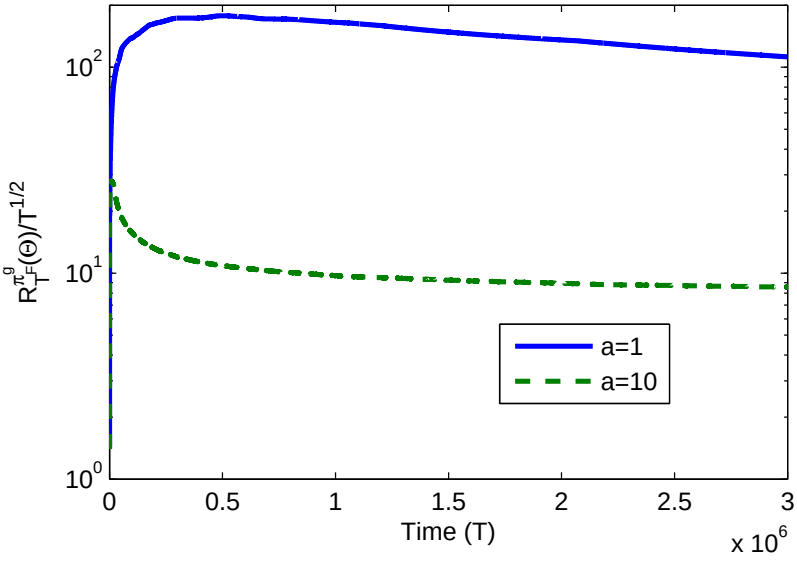


Fig. 7. The convergence of the regret ($M = 2,\ N = 9,\ \Theta = [0.1,\ 0.2,\ \cdots,\ 0.9]$).

events.

$$\begin{aligned} A_r &\triangleq \cap_{i\le n\le N}\{\max_{\delta c^{r-1}\le s} |\tilde{\theta}_{\sigma(n)\ (s \text{ samples})} - \theta_{\sigma(n)}| \le \alpha_1\}, \\ B_r &\triangleq \{\tilde{\theta}_{\sigma(i)\ (j \text{ samples})} \ge \theta_{\sigma(i)} - \alpha_1 \text{ or } I(\tilde{\theta}_{\sigma(i)\ (j \text{ samples})}, \theta_{\sigma(i)} - \alpha_1) \le \log(s_m - 1)/j \\ &\quad \text{for all } 1 \le j \le bm,\ c^{r-1} \le m \le c^{r+1},\ \text{and } s_m > m\}. \end{aligned} \tag{12}$$

By (10), we have $\Pr(\bar{A}_r) = o(c^{-r})$. Consider the following event:

$$C_r \triangleq \{\tilde{\theta}_{\sigma(i)\ (j \text{ samples})} \ge \theta_{\sigma(i)} - \alpha_1 \text{ or } I(\tilde{\theta}_{\sigma(i)\ (j \text{ samples})}, \theta_{\sigma(i)} - \alpha_1) \le \log(m)/j \text{ for all } 1 \le j \le bm,\ c^{r-1} \le m \le c^{r+1}\}. \tag{13}$$

We have that $B_r \supset C_r$. From Lemma $1-(i)$ in [4], $\Pr(\bar{C}_r) = o(c^{-r})$. We thus have $\Pr(\bar{B}_r) = o(c^{-r})$.

Consider the $s$th update where $c^{r-1} \le s - 1 < c^{r+1}$. When the round-robin candidate $r_t = \sigma(i)$, we show that on the event $A_r \cap B_r$, $\sigma(i)$ must be selected as the $i$th best. It is sufficient to focus on the nontrivial case that $\theta_{l_t} < \theta_{\sigma(i)}$. Since $\tau_{l_t,t} \ge (s-1)b$, on $A_r$, we have $\tilde{\theta}_{l_t}(t) < \theta_{\sigma(i)} - \alpha_1$. We also have, on $A_r \cap B_r$,

$$\tilde{\theta}_{\sigma(i)}(t) \ge \theta_{\sigma(i)} - \alpha_1 \text{ or } I(\tilde{\theta}_{\sigma(i)}(t), \theta_{\sigma(i)} - \alpha_1) \le \log(t-1)/\tau_{\sigma(i),t}. \tag{14}$$

Channel $\sigma(i)$ is thus selected as the $i$th best on $A_r \cap B_r$. Since $(1-c^{-1})/N > b$, for any $c^r \le s-1 \le c^{r+1}$, there exists an $r_0$ such that on $A_r \cap B_r$, $\tau_{\sigma(i),t} \ge (1/N)(s - c^{r-1} - 2N) > bs$ for all $r > r_0$. It thus

follows that on $A_r \cap B_r$, for any $c^r \leq s-1 \leq c^{r+1}$, we have $\tau_{\sigma(i),t} > (s-1)b$, and $\sigma(i)$ is thus the leader. We have, for all $r > r_0$,

$$\Pr(\text{at the } s\text{th update, } c^{r-1} \leq s-1 < c^{r+1}, l_t \neq \sigma(i)) \leq \Pr(\bar{A}_r) + \Pr(\bar{B}_r) = o(c^{-r}). \tag{15}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[N_3(K)] &= \mathbb{E}[|\{1 \leq k \leq K : \ k \in \mathcal{D}(K), \ l_t \neq \sigma(i)\}|] \\
&\leq \Sigma_{s=1}^{T} \Pr(\text{at the } s\text{th update, } l_t \neq \sigma(i)) \\
&\leq 1 + \Sigma_{r=0}^{\lceil \log_c T \rceil} \Sigma_{c^r \leq s-1 \leq c^{r+1}} \Pr(\text{at the } s\text{th update, } l_t \neq \sigma(i)) \\
&= 1 + \Sigma_{r=0}^{\lceil \log_c T \rceil} o(1) \\
&= o(\log T).
\end{aligned} \tag{16}$$

From (9), (11), (16), we arrive at Lemma 2.

## REFERENCES


[1] L. Lai, H. Jiang and H. Vincent Poor, "Medium Access in Cognitive Radio Networks: A Competitive Multi-armed Bandit Framework," in *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008.

[2] K. Liu and Q. Zhao, "Decentralized Multi-Armed Bandit with Distributed Multiple Players," in *Proc. of Information Theory and Applications Workshop (ITA)*, January, 2010.

[3] A. Anandkumar, N. Michael, and A.K. Tang, "Opportunistic Spectrum Access with Multiple Players: Learning under Competition," in *Proc. of IEEE INFOCOM*, Mar. 2010.

[4] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4C22, 1985.

[5] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: IID rewards," *IEEE Tran. on Auto. Control*, vol. 32, no. 11, pp. 968C976, 1987.

[6] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, Vol. 47, pp. 235-256, 2002.

[7] Y. Gai, B. Krishnamachari, and R. Jain, "Learning Multiplayer Channel Allocations in Cognitive Radio Networks: A Combinatorial Multi-Armed Bandit Formulation," in *Proc. of IEEE DySPAN*, 2010.

[8] B. C. Levy, "Principles of Signal Detection and Parameter Estimation," Springer, July, 2008.